CLAIMS

1. A method for extracting information from a natural language text corpus based on a natural language query, comprising the steps of:

analyzing said natural language text corpus with respect to surface structure of word tokens and surface syntactic roles of constituents;

indexing and storing the analyzed natural language text corpus;

analyzing a natural language query with respect to surface structure of word tokens and surface syntactic roles of constituents;

creating one or more surface variants of the analyzed natural language query, said one or more surface variants being equivalent to said natural language query with respect to lexical meaning of word tokens and surface syntactic roles of constituents;

comparing said one or more surface variants and said analyzed natural language query with the indexed and stored analyzed natural language text corpus; and

extracting from said indexed and stored analyzed natural language text corpus, each portion of text comprising a string of word tokens that matches any one of said surface variants or said analyzed natural language query.

2. The method according to claim 1, wherein, in the step of creating, said surface syntactic roles of constituents are head and modifier roles, and grammatical relations.

3. The method according to claim 1, wherein, in the step of extracting, a string of word tokens in said indexed and stored analyzed natural language text corpus matches one of said surface variants or said analyzed natural language query if it comprises the head words of

phrases bearing the grammatical relations of subject, object, and lexical main verb in said one of said surface variants or said analyzed natural language query in the same linear order as in said one of said surface variants or said analyzed natural language query.

4. The method according to claim 1, wherein, in the step of analyzing a natural language query, said natural language query is analyzed in the same manner as said natural language text corpus is analyzed in the step of analyzing said natural language text corpus.

5. The method according to claim 1, wherein the step of analyzing a natural language text corpus comprises the steps of:

determining a morpho-syntactic description for each word token of said natural language text corpus;

locating phrases in said natural language text corpus;

determining a phrase type for each of said phrases; and

locating clauses in said natural language text corpus,

and wherein the step of analyzing a natural language query comprises the steps of:

determining a morpho-syntactic description for each word token of said natural language query; and

locating phrases in said natural language query;

determining a phrase type for each of said phrases; and

locating clauses in said natural language query.

6. The method according to claim 5, wherein the step of indexing and storing comprises the steps of:

providing, for each word token of said natural language text corpus with, a unique word token location identifier;

storing information regarding the location of each word token of said natural language text corpus, based on said unique word token location identifiers;

storing, for each phrase type, information regarding the location of each phrase of this type in said natural language text corpus, based on said unique word token location identifiers; and

storing information regarding the location of each clause in said natural language text corpus, based on said unique word token location identifiers.

7. The method according to claim 6, wherein each word token is associated with a word type, and wherein the step of storing information regarding the location of each word token comprises the steps of:

storing each word type of said natural language text corpus; and

storing, for each word token, its unique word token location identifier logically linked to the stored associated word type.

8. The method according to claim 7, wherein the step of storing information regarding the locations of phrases comprises the steps of:

providing, for each phrase of said natural language text corpus, a unique phrase location identifier identifying the word tokens spanned by the phrase;

storing each phrase type of said natural language text corpus; and

storing, for each phrase, its unique phrase location identifier logically linked to the stored associated phrase type.

9. The method according to claim 8, wherein the step of storing information regarding the locations of clauses comprises the steps of:

providing, for each clause of said natural language text corpus, a unique clause location identifier identifying the word tokens and phrases spanned by the clause;

storing, for each clause, its unique clause location identifier.

10. The method according to claim 9, further comprising the steps of:

locating sentences in said natural language text corpus; and

providing, for each sentence of said natural language text corpus, a unique sentence location identifier identifying the word tokens, phrases and clauses spanned by the sentence;

storing, for each sentence, its unique sentence location identifier.

11. The method according to claim 10, further comprising the steps of:

locating paragraphs in said natural language text corpus;

providing, for each paragraph of said natural language text corpus, a unique paragraph location identifier identifying the word tokens, phrases, clauses and sentences spanned by the paragraph;

storing, for each paragraph, its unique paragraph location identifier.

12. The method according to claim 11, further comprising the steps of:

locating documents in said natural language text corpus;

providing, for each document of said natural language text corpus, a unique document location identifier identifying the word tokens, phrases, clauses, sentences and paragraphs spanned by the document;

storing, for each document, its unique document location identifier.

13. The method according to claim 1, wherein, in the step of extracting, a portion of text that is extracted is either the matching string of word tokens, a clause comprising the matching string of word tokens, a sentence comprising the matching string of word tokens, a paragraph comprising the matching string of word tokens, or a document comprising the matching string of word tokens.

14. The method according to claim 1, further comprising the step of:

organizing the extracted information according to degree of correspondence with the query with respect to lexical meaning of word tokens and surface syntactic roles of constituents, such that a constituent in a portion of text having the same lemma as the equivalent constituent of the query is considered to have a higher degree of correspondence than a constituent in a portion of text being a synonym to the equivalent constituent of the query.

15. The method according to claim 1, further comprising the step of:

organizing the extracted information such that said portions of text are grouped according to sameness of grammatical subject, grammatical object, and lexical main verb.

16. A system for extracting information from a natural language text corpus based on a natural language query, comprising:

a text analysis unit for analyzing a natural language text corpus and a natural language query with

respect to surface structure of word tokens and surface syntactic roles of constituents;

storage means operatively connected to said text analysis unit, for storing the analyzed natural language text corpus;

an indexer, operatively connected to said storage means, for indexing the analyzed natural language text corpus;

an index, operatively connected to said indexer, for storing said indexed analyzed natural language text corpus;

a query manager, operatively connected to said text analysis unit, comprising means for creating surface variants of said natural language query, said surface variants being equivalent to said natural language query with respect to lexical meaning of word tokens and surface syntactic roles of constituents, and means for comparing said surface variants and said analyzed natural language query with the indexed analyzed natural language text corpus in said index; and

a result manager operatively connected to said index, for extracting, from said indexed and stored analyzed natural language text corpus, each portion of text comprising a string of word tokens that matches any one of said surface variants or said analyzed natural language query.

17. The system according to claim 16, wherein a string of word tokens in said indexed and stored analyzed natural language text corpus matches one of said surface variants or said analyzed natural language query if it comprises the head words of phrases bearing the grammatical relations of subject, object, and lexical main verb in said one of said surface variants or said analyzed natural language query in the same linear order as in said one of said surface variants or said analyzed natural language query.

18. The system according to claim 16, wherein said index comprises multiple indexes based on a hierarchy of text units that are related by inclusion.

19. A computer readable medium having computer-executable instructions for a general-purpose computer to perform the steps recited in claim 1.

20. A computer program comprising computer-executable instructions for performing the steps recited in claim 1.